DATA 101 - Final

Beau Churchill

2025-08-09

Topic:

Investigating patterns in police officer deaths over time and their potential relationship with environmental factors such as temperature and seasonal trends.

Dataset:

For this research, two datasets are utilized to explore potential relationships between environmental factors and police officer fatalities. The first dataset, police_deaths.csv, provides comprehensive information about officer fatalities, including the date of death (End_Of_Watch), cause of death, age, department, state, and other relevant details. This data allows for analysis of trends in fatalities over time, causes, and demographic or geographic patterns.

Source: Kaggle

The second dataset, average_temperature.csv, contains historical records of average yearly temperatures, with variables for the year and the corresponding average temperature. By combining these datasets, the study aims to investigate whether fluctuations in average temperature across years correspond with changes in the number of police deaths, offering insight into how environmental conditions might impact officer safety.

Source: Kaggle

Introduction:

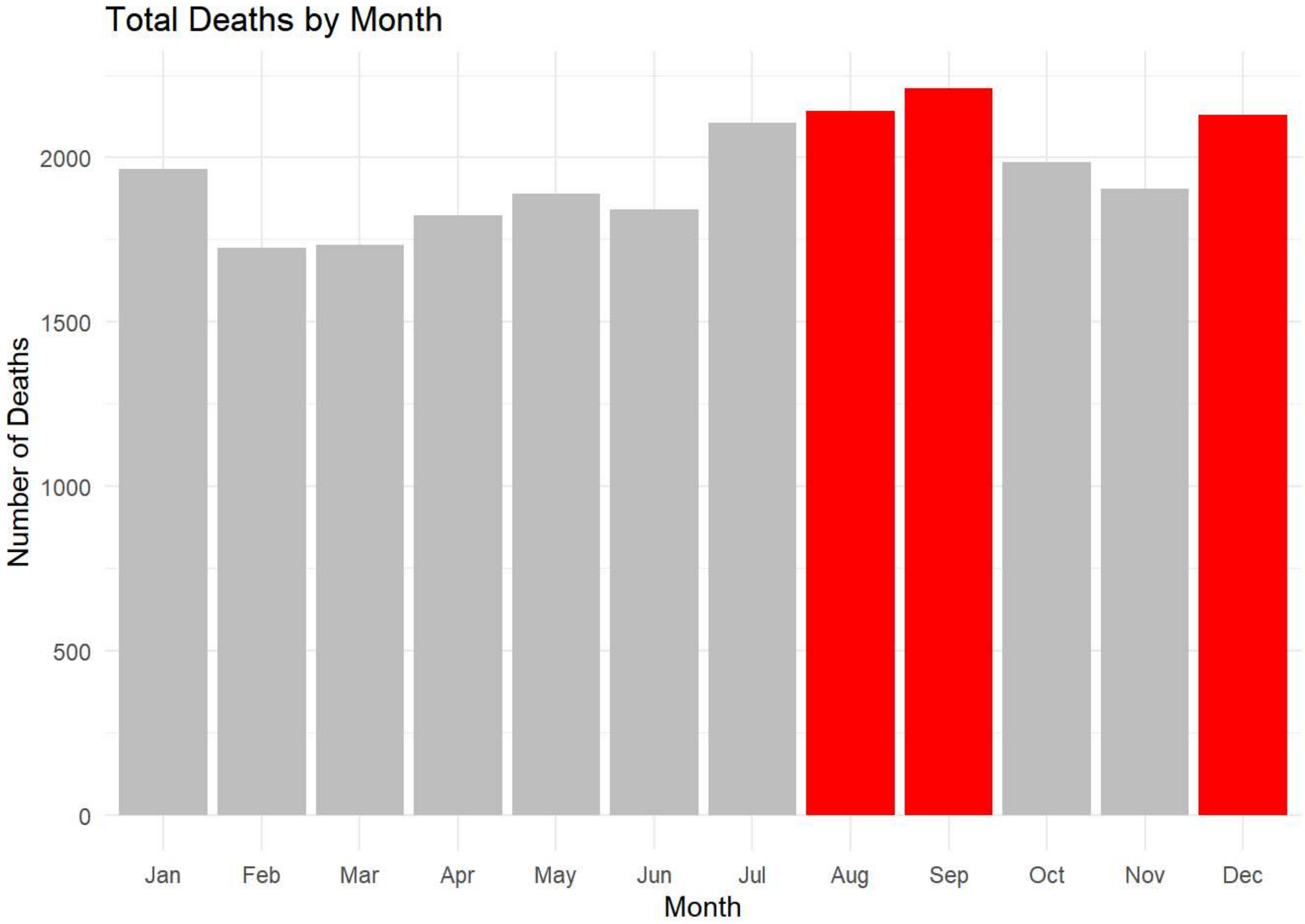
This project seeks to investigate whether there is a relationship between police officer fatalities and average yearly temperature. Specifically, the question addressed is: Does the average temperature in a given year correlate with the number of officer deaths in that year? The null hypothesis (H_0) states that there is no significant relationship between yearly average temperature and the number of officer fatalities. The alternative hypothesis (H_1) predicts that higher average temperatures—particularly extreme heat—are associated with an increase in officer deaths.

To answer this, two datasets are used. The primary dataset, police_deaths.csv, contains detailed records of law enforcement fatalities including the date of death (End_Of_Watch), cause, age, department, and location. For this analysis, the End_Of_Watch date is the most critical variable, as it allows aggregation of deaths by year and month. The second dataset, average_temperature.csv, contains yearly U.S. average temperature data, with columns for the Year and the average Temperature measured in degrees Fahrenheit. By combining these datasets on the Year variable, the project aims to explore any temporal patterns or correlations between environmental temperature and officer fatalities.

Data Analysis

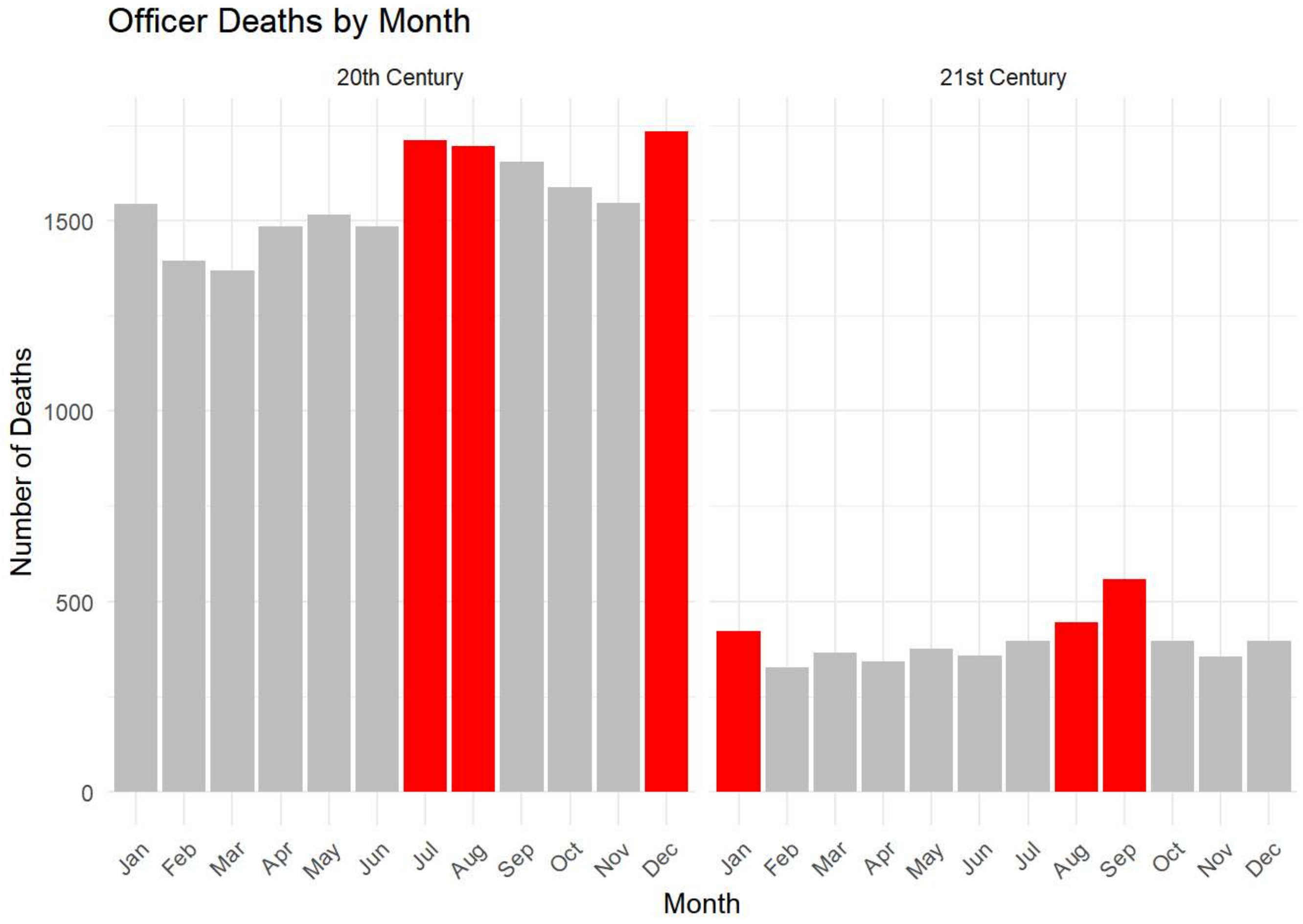
The data analysis begins with cleaning and preparing the datasets for comparison. The police deaths data is parsed to extract the year and month of each fatality, allowing aggregation of deaths by month and year. Originally, the observations for police deaths started in the 1700s, which I cleaned to align with the weather dataset that starts in 1900. An exploratory analysis identified the three months with the highest numbers of officer deaths across all years, which are then used to focus the analysis on the periods with the greatest risk. Separately, the temperature data is aggregated to obtain average yearly temperatures. These datasets are then merged by year to align fatalities with temperature data. Correlation analysis is performed to measure the strength and direction of any association between average yearly temperature and officer deaths during the top three months. To visualize this relationship, scatter plots with fitted linear regression lines are generated. This approach helps in understanding whether environmental temperature is a potential factor influencing fatality rates in law enforcement, with the visualization providing a clear depiction of trends and the correlation coefficient quantifying the strength of the relationship.

```r
# Create the monthly counts data frame
monthly_counts <- data %>%
  mutate(
    date = parse_date_time(End_Of_Watch, orders = c("ymd", "mdy", "dmy")),
    month_of_death = month(date, label = TRUE)
  ) %>%
  count(month_of_death) %>%
  arrange(desc(n))

# top three months
top_three <- monthly_counts %>%
  slice_head(n = 3) %>%
  pull(month_of_death)

# Create bar chart
ggplot(monthly_counts, aes(x = month_of_death, y = n, fill = month_of_death %in% top_three)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(
    values = c("gray", "red"),
    labels = c("Other Months", "Top 3 Months")) +
  labs(
    title = "Total Deaths by Month",
    x = "Month",
    y = "Number of Deaths"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

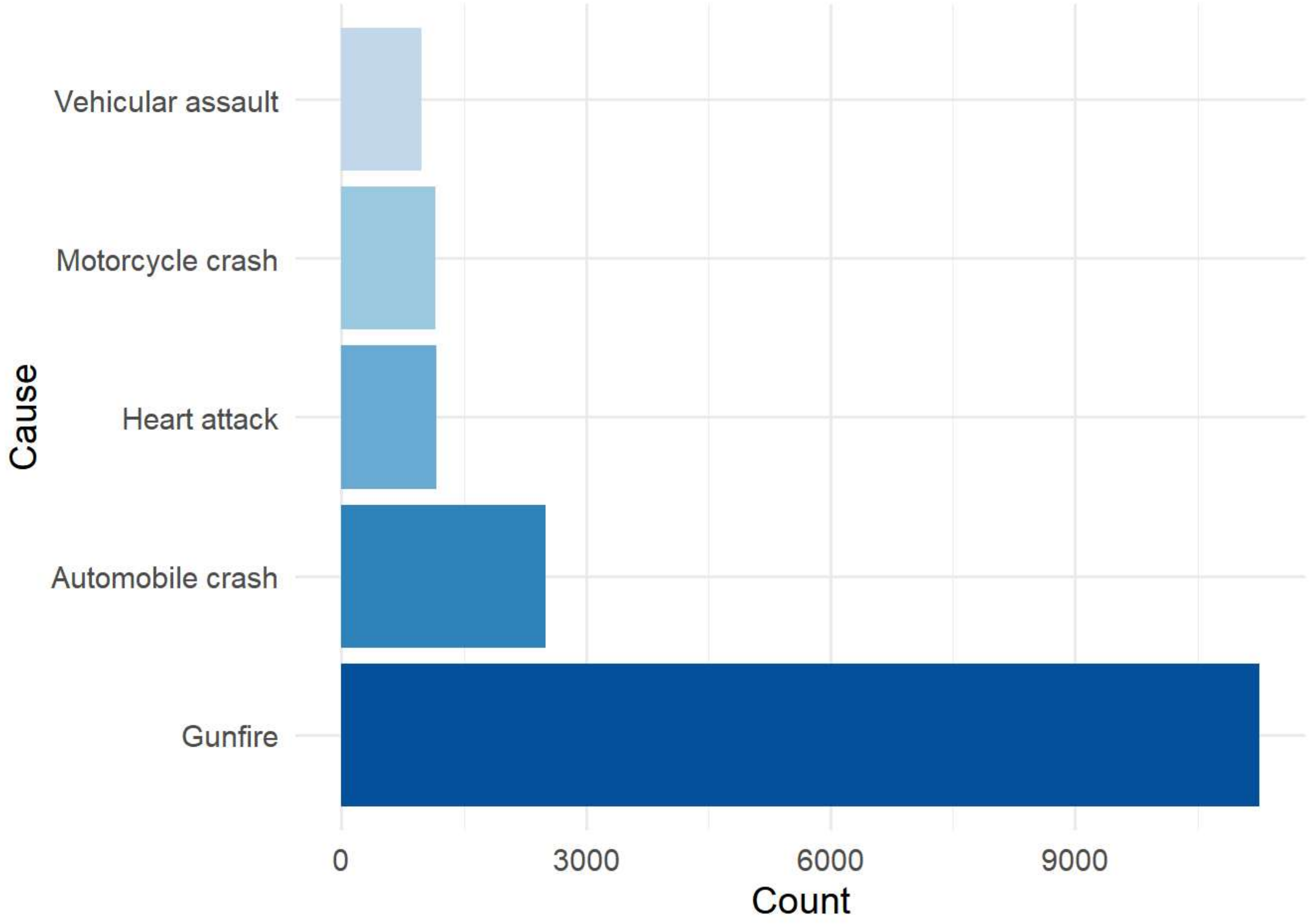

Total Deaths by Month
Number of Deaths
2000
1500
1000
500
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

```
# ***Exploratory Analysis***

# Parse dates and extract year, century, and month
data_century_month <- data %>%
  mutate(
    date = parse_date_time(End_Of_Watch, orders = c("ymd", "mdy", "dmy")),
    year = year(date),
    century = case_when(
      year >= 1800 & year < 1900 ~ "19th Century",
      year >= 1900 & year < 2000 ~ "20th Century",
      year >= 2000 & year < 2100 ~ "21st Century",
      TRUE ~ NA_character_
    ),
    month_of_death = month(date, label = TRUE, abbr = TRUE)
  ) %>%
  filter(!is.na(century)) %>%
  group_by(century, month_of_death) %>%
  summarise(deaths = n(), .groups = "drop") %>%
  mutate(
    month_of_death = as.character(month_of_death),
    century = as.character(century)
  )

# Find top 3 months per century
top_months_per_century <- data_century_month %>%
  group_by(century) %>%
  slice_max(order_by = deaths, n = 3) %>%
  ungroup() %>%
  mutate(
    month_of_death = as.character(month_of_death),
    century = as.character(century)
  )

# Join to flag top 3 months; this join may create duplicate deaths columns
data_century_month <- data_century_month %>%
  left_join(
    top_months_per_century %>%
      mutate(is_top3 = TRUE),
    by = c("century", "month_of_death")
  ) %>%
  mutate(is_top3 = if_else(is.na(is_top3), FALSE, TRUE))

# Fix duplicate deaths columns if any (rename deaths.x and remove deaths.y)
if ("deaths.x" %in% names(data_century_month)) {
  data_century_month <- data_century_month %>%
    rename(deaths = deaths.x) %>%
    select(-deaths.y)
}

# Re-factor month_of_death for correct month ordering in the plot
data_century_month$month_of_death <- factor(data_century_month$month_of_death, levels = month.abb)
```

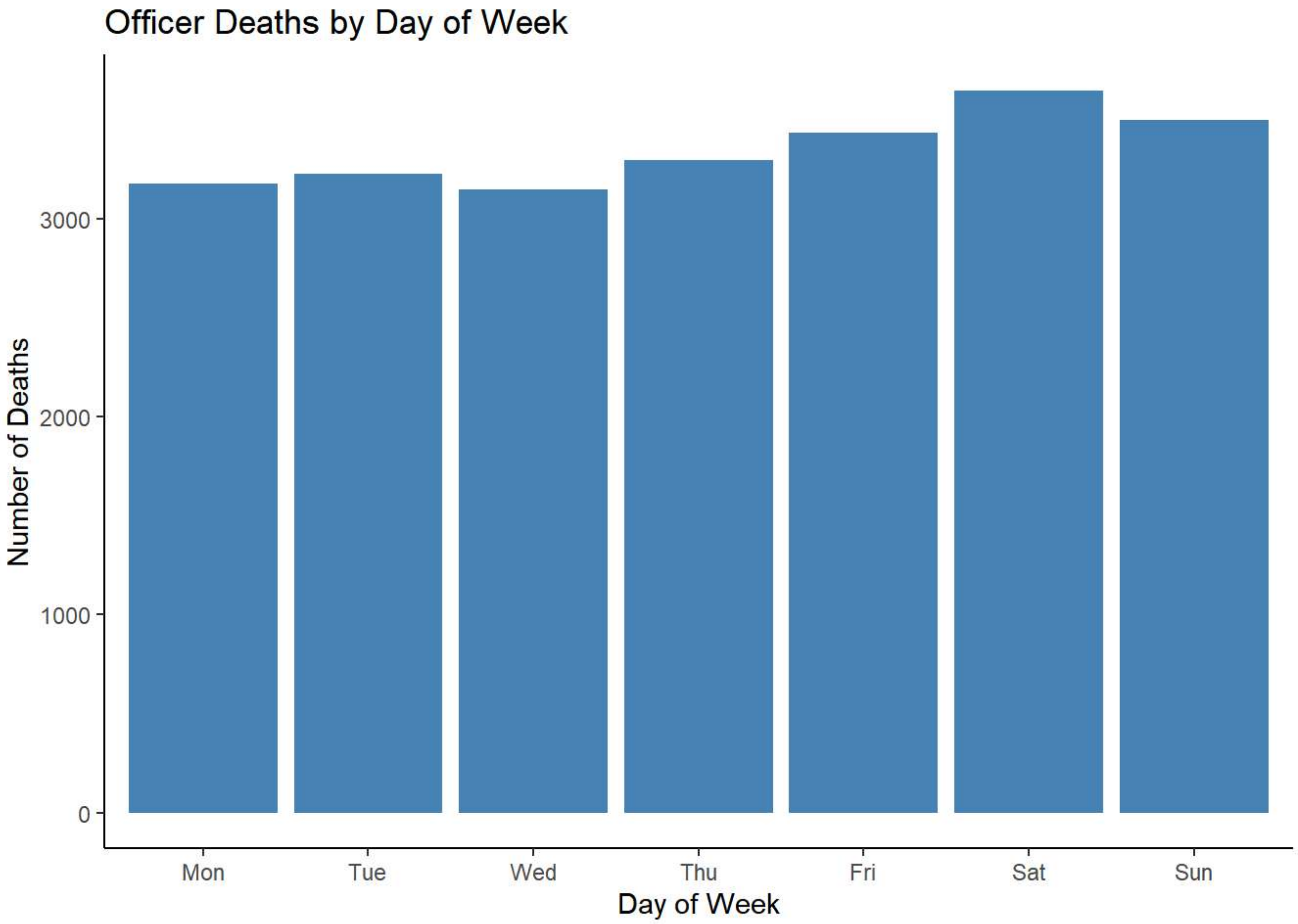
```
ggplot(data_century_month, aes(x = month_of_death, y = deaths, fill = is_top3)) +
  geom_col() +
  facet_wrap(~ century) +
  scale_x_discrete(drop = FALSE) +   # show all months even if zero deaths
  scale_fill_manual(values = c("gray", "red"), labels = c("Other Months", "Top 3 Months")) +
  labs(
    title = "Officer Deaths by Month",
    x = "Month",
    y = "Number of Deaths"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```



```
# ***Exploratory Analysis***

# Create top 5 causes dataset
top_5_causes_data <- data %>%
  count(Cause, name = "count") %>%
  arrange(desc(count)) %>%
  slice(1:5) %>%
  mutate(Cause = fct_reorder(Cause, count, .desc = TRUE))

# Define colors for my theme
police_colors <- c("#08519c", "#3182bd", "#6baed6", "#9ecae1", "#c6dbef")

# Plot for top 5 causes of deaths
ggplot(top_5_causes_data, aes(x = Cause, y = count, fill = Cause)) +
  geom_col() +
  scale_fill_manual(values = police_colors) +
  coord_flip() +
  labs(x = "Cause", y = "Count") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```



```
# Plot for deaths by day of week

# Define the full names and abbreviations for days of the week
day_levels <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
day_abbrev <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

# Clean up the Day_Of_Week column
data <- data %>%
  mutate(
    Day_Of_Week = str_trim(Day_Of_Week),

# Remove whitespace
Day_Of_Week = str_to_title(Day_Of_Week)) %>%

# Capitalize first letter
  mutate(Day_Of_Week = factor(Day_Of_Week, levels = day_levels)) %>%
  filter(!is.na(Day_Of_Week))

# Remove rows where factor failed

ggplot(data, aes(x = Day_Of_Week)) +
  geom_bar(fill = "steelblue") +
  scale_x_discrete(labels = day_abbrev) +

# Abbreviate day names here
  labs(
    title = "Officer Deaths by Day of Week",
    x = "Day of Week",
    y = "Number of Deaths"
  ) +
  theme_classic()
```

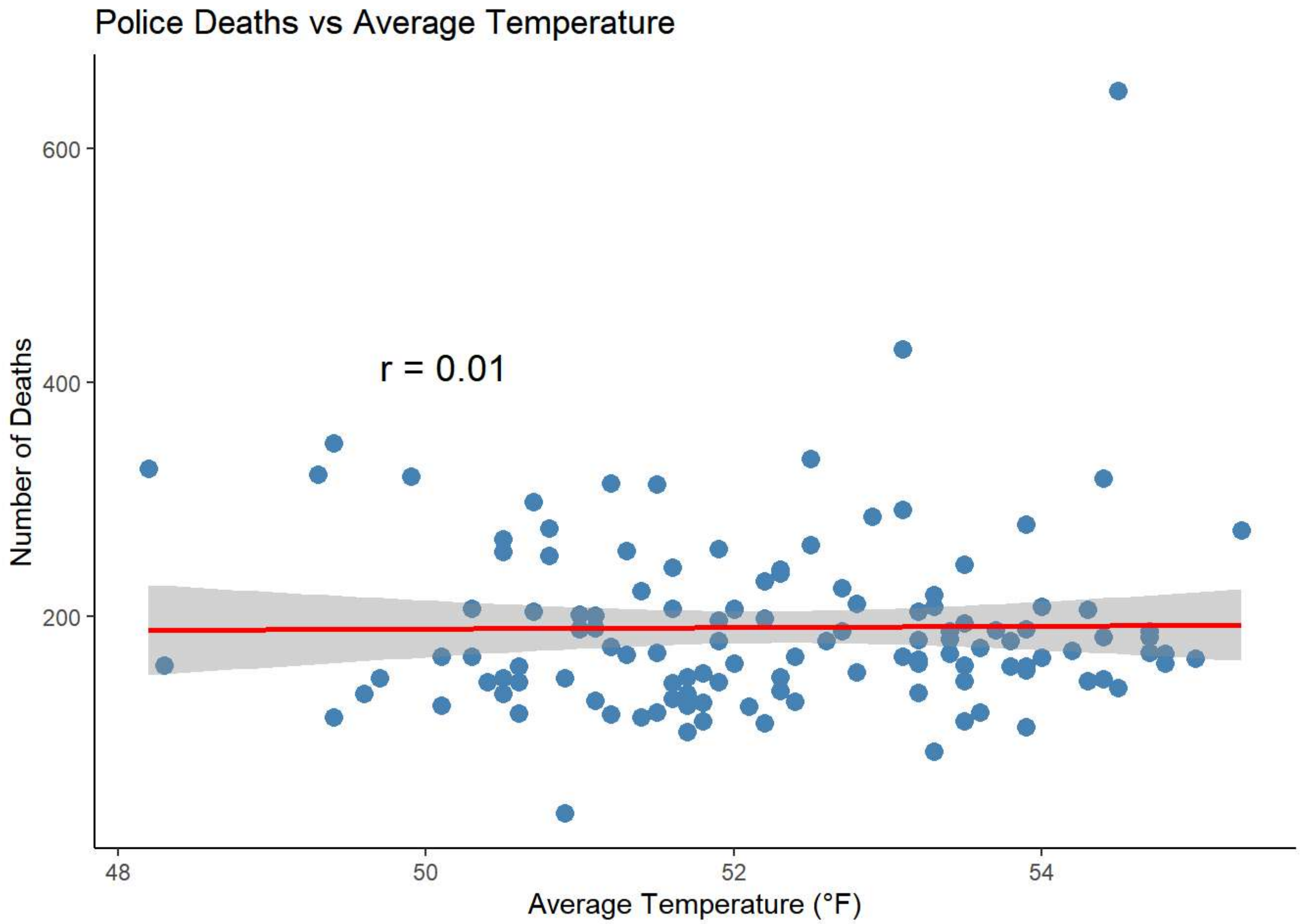


```
# Plot to see if there is correlation between avg yearly temps to avg yearly deaths.

# Extract year and count deaths per year
police_yearly <- data %>%
  mutate(Year = year(parse_date_time(End_Of_Watch, orders = c("ymd", "mdy", "dmy")))) %>%
  filter(!is.na(Year)) %>%
  count(Year, name = "Deaths")

# Summarize Year and Temperature column
temp_yearly <- average_temperature %>%
  group_by(Year) %>%
  summarise(Avg_Temperature = mean(Temperature, na.rm = TRUE))

# Join by Year
combined <- police_yearly %>%
  inner_join(temp_yearly, by = "Year")

print(combined)
```

```
## # A tibble: 123 × 3
##     Year Deaths Avg_Temperature
##    <dbl>  <int>           <dbl>
##  1  1900    105            53.9
##  2  1901    110            53.5
##  3  1902    123            52.1
##  4  1903    117            50.6
##  5  1904    110            51.8
##  6  1905    101            51.7
##  7  1906    116            51.2
##  8  1907    127            52.4
##  9  1908    166            50.1
## 10  1909    126            51.8
## # ℹ 113 more rows
```

```
# Scatter plot - Police Deaths vs. Avg Temps

# Calculate correlation coefficient
corr_coef <- cor(combined$Avg_Temperature, combined$Deaths, use = "complete.obs")

# Choose an x position near the middle or max of your data range
x_pos <- median(combined$Avg_Temperature, na.rm = TRUE)

# Calculate linear regression line at x_pos
lm_fit <- lm(Deaths ~ Avg_Temperature, data = combined)
y_pos <- predict(lm_fit, newdata = data.frame(Avg_Temperature = x_pos))



# Scatter Plot
ggplot(combined, aes(x = Avg_Temperature, y = Deaths)) +
  geom_point(color = "steelblue", size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  annotate(
    "text",
    x = x_pos,
    y = y_pos,
    label = paste0("r = ", round(corr_coef, 2)),
    color = "black",
    vjust = -9,
    hjust = 3,
    size = 5
  ) +
  labs(
    title = "Police Deaths vs Average Temperature",
    x = "Average Temperature (°F)",
    y = "Number of Deaths"
  ) +
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion and Future Directions

The null hypothesis for this project proposed that average yearly temperature has no significant effect on the number of police officer deaths, while the alternative hypothesis predicted that higher heat would increase fatalities. Based on the correlation coefficient — which was close to zero — there was insufficient evidence to reject the null hypothesis. This means that, within the limits of the available data, temperature does not appear to have a measurable linear effect on police officer deaths.

The analysis revealed that there is little to no significant correlation between average yearly temperature and the number of police officer fatalities, including during the three months with the highest death counts. The correlation coefficient was close to zero, indicating no clear linear relationship between temperature and fatality rates. This suggests that factors other than annual average temperature are likely more influential in determining police deaths, at least within the scope of the data analyzed.

However, this study's reliance on yearly average temperature data may mask more nuanced relationships between weather and fatalities. Because weather conditions vary significantly within a year, analyzing monthly temperature data in conjunction with monthly fatality counts would provide a more detailed understanding of potential seasonal or short-term weather effects on police deaths. Future research could expand on this by incorporating monthly or even daily weather data, as well as exploring other environmental or situational variables such as precipitation, extreme weather events, or location-specific climate patterns. Additionally, examining other risk factors and contextual variables, such as changes in law enforcement practices or crime rates over time, could further clarify the complex dynamics influencing officer fatalities.

Sources:

Police Officer Deaths in the U.S. 5 Nov. 2016, www.kaggle.com/datasets/fivethirtyeight/police-officer-deaths-in-the-us.

National Centers for Environmental Information (NCEI). www.ncei.noaa.gov.

Average Temperature From 1900 to 2023. 25 Nov. 2023, www.kaggle.com/datasets/giabchnguyn/average-temperature-from-1900-to-2023.

W3Schools.com. www.w3schools.com/r.